

Article

Reliability-Controlled Head Adaptation for Gait Prediction with Wearable Devices Under Unreliable Calibration

Zhiyuan Zhou ¹ and Kewei Liang ^{2,*} 

¹ Faculty of Science and Technology, Beijing Normal-Hong Kong Baptist University (BNBU), Zhuhai 519087, China; t330033058@mail.bnpu.edu.cn

² School of Mathematical Sciences, Zhejiang University, Hangzhou 310058, China

* Correspondence: matlkw@zju.edu.cn

Abstract

When a wearable gait predictor is personalized to a new user, a short calibration session may contain missing channels, sensor-placement changes, electrode shift, noise, fatigue, or other non-ideal effects. Treating all calibration samples as equally trustworthy can overpersonalize the prediction head and induce harmful drift away from the source model. We study calibration quality as a control signal for new-user personalization and propose Dynamic Trust-Region Head Adaptation (DTR-HA), a head-only adaptation rule that combines reliability-weighted calibration loss with a reliability-scaled source-head anchor. The temporal encoder is frozen, and lower reliability strengthens a soft source-head penalty. The Bilateral Lower-Limb Neuromechanical Signals dataset (BLISS) defines the target early gait-phase prediction task with 21-subject leave-one-subject-out evaluation; complete-bout calibration/test separation; and 0, 100, and 200 ms horizons. K2MUSE supplies real non-ideal calibration conditions for mechanism-level stress testing. In K2MUSE 75% bad-calibration tests with a held-out blind context-based reliability scorer, DTR-HA reduced empirical risk by 20–23% relative to plain head adaptation, improved macro-F1 by 0.018–0.022, and achieved 25/28 paired subject–condition wins with fewer negative-adaptation cases. In a secondary BLISS calibration-pollution check, DTR-HA kept macro-F1 within 0.003 of plain head adaptation while reducing calibration-induced head drift across all horizons. These results support calibration quality as an adaptation-stage control signal for stable, lightweight wearable gait personalization under unreliable calibration.

Keywords: wearable sensing; lower-limb gait prediction; reliability-aware adaptation; few-shot personalization; calibration quality; sensor degradation; new-user adaptation



Academic Editor: Rui Araújo

Received: 7 July 2026

Revised: 27 July 2026

Accepted: 30 July 2026

Published: 2 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Wearable lower-limb sensing is an applied motion-prediction problem for gait monitoring, lower-limb assistive devices, prostheses, exoskeletons, and rehabilitation robots. A sensing front end that estimates gait phase or locomotor intent can inform when a downstream system should prepare assistance, reduce assistance, request recalibration, or remain in a conservative state [1–5]. In these systems, prediction errors are not merely offline classification mistakes: unreliable personalization can move the sensing front end toward a decision boundary that is inconsistent with the user’s true locomotor state. Overconfident or mistimed predictions may destabilize assistance timing, and an uncertainty-aware fallback can be preferable to forcing a low-quality prediction.

Wearable gait-prediction models are commonly evaluated as recognition or forecasting models. New-user personalization adds a second decision: after a short calibration session is collected, should the source model be adapted, and how strongly should those data influence the update? Calibration data in wearable sensing can be sparse, subject-specific, and non-ideal. A model is often trained on a group of subjects and then used on a new individual whose gait dynamics, neuromuscular activation, impairment level, walking speed, and sensor placement can differ substantially from those of the training set. Few-shot personalization can reduce this mismatch, but it introduces another practical assumption: the calibration data provided by the new user are reliable. In lower-limb wearable sensing, this assumption is fragile. Surface electromyography (EMG) patterns can vary with arm position, electrodes can shift, contact quality can degrade, fatigue can alter activation patterns, channels can be missing, and inter-day acquisition differences can move the calibration distribution away from the source distribution [6–8]. Thus, this study centers on adaptation-stage reliability rather than increasing backbone capacity.

We ask whether calibration reliability can regulate prediction-head adaptation and reduce negative adaptation under polluted calibration. Negative adaptation means that personalization lowers macro-F1 relative to the source-only model for the same held-out user and test set. The method studied here, Dynamic Trust-Region Head Adaptation (DTR-HA), is a reliability-controlled source-anchoring rule for head-only new-user adaptation. For each held-out user, a source temporal model is trained on the training subjects. During personalization, the temporal encoder is frozen and only the task-specific prediction head is adapted. DTR-HA uses reliability at two levels: window-level scores weight the calibration loss, and calibration-set reliability scales a soft penalty around the source head. Lower reliability therefore increases the penalty on head movement instead of imposing a hard displacement limit.

Public datasets are used deliberately to evaluate this engineering question under reproducible conditions. The Datasets for Bilateral Lower-Limb Neuromechanical Signals in Able-Bodied and Impaired Individuals with Wearable and Ambient Sensors (BLISS) [9] establish the main early gait-phase prediction protocol: full 21-subject leave-one-subject-out (LOSO) evaluation; future prediction at 0, 100, and 200 ms; complete-bout calibration/test separation; and few-shot head adaptation. K2MUSE [8] provides real non-ideal acquisition variability, including fatigue- and shift-related calibration conditions. An additional BLISS target-pipeline stress test checks whether the same drift-control behavior appears when only held-out BLISS calibration bouts are polluted. Existing studies of wearable gait-recognition technology often emphasize model architecture or recognition accuracy; the present work instead focuses on calibration quality as an adaptation-stage variable during new-user personalization.

This study's main contributions to the literature are as follows:

1. **Method:** DTR-HA is a reliability-scaled source-anchoring rule that combines window-level reliability weighting with calibration-set control of the prediction-head penalty.
2. **Evaluation design:** This work uses a leakage-safe BLISS protocol for 21-subject LOSO early gait-phase prediction with complete-bout calibration/test separation. This is an evaluation design for the present study, not a new dataset or a separate method.
3. **Evidence:** K2MUSE provides real non-ideal calibration stress evidence, while the additional BLISS calibration-pollution check tests drift control inside the target gait-prediction pipeline.

2. Related Work

2.1. Wearable Gait-Prediction Technology and Public Benchmark Evaluation

Lower-limb wearable interfaces commonly combine neuromuscular and mechanical signals, including EMG, inertial measurements, kinematics, and kinetics, to support locomotion recognition, gait-event detection, and exoskeleton or prosthesis front ends [1–3,10–13]. Reviews of active lower-limb devices emphasize that controller performance depends on reliable sensing and state estimation [4,14,15]. Applied studies span real-time gait-phase estimation, prosthetic walking-speed estimation, joint-moment prediction, and subject-specific gait-phase feature selection [16–19]. These studies establish the engineering value of wearable motion-prediction technology, but they do not make calibration quality the variable that governs a new-user update. Deep temporal models such as convolutional and recurrent networks are attractive for time-series prediction [20–22], but increasing model capacity alone does not resolve unreliable calibration.

Early gait-phase prediction is harder than current-state recognition because the target may change after the end of the observation window. In assistive devices, this is also the more relevant setting: a controller often needs time to prepare assistance before a phase transition is fully visible. A protocol that evaluates only current-phase classification can therefore overestimate deployment usefulness. This paper treats current-phase recognition as the 0 ms horizon and evaluates future-phase prediction at 100 and 200 ms using the same held-out user and complete-bout split.

2.2. New-User Personalization in Wearable Sensing

Subject variability is a major barrier in lower-limb wearable interfaces. Human-in-the-loop optimization and personalization studies show that user-specific adaptation can improve wearable devices' assistance and recognition performance [23–25]. Gait-phase studies likewise report user-dependent feature and model behavior [19]. In practice, however, user-specific calibration is often collected quickly and under imperfect sensing conditions. Meta-learning methods such as model-agnostic meta-learning (MAML) learn an initialization for rapid gradient-based adaptation, whereas prototypical networks learn an episodic metric space for few-shot classification [26,27]. Conditional modulation and parameter-efficient adapters offer other ways to limit task-specific updates [28,29]. These families address how a model should adapt efficiently; the present study asks how strongly a completed calibration session should be trusted when it may be polluted.

The relevant distinction is between using calibration data after it is collected and deciding how strongly the model should trust it. Existing studies of wearable gait-prediction technology and personalization establish the sensing and adaptation context, but they typically do not use calibration quality to scale a source-centered penalty during new-user head adaptation.

2.3. Sensor Degradation and Unreliable Calibration

Wearable sensors are vulnerable to missingness, outliers, contact variation, fatigue, and placement changes. EMG is especially sensitive to arm-position variation, electrode shift, baseline noise, motion artifacts, and physiological variability [6,7,30,31]. K2MUSE is relevant because it explicitly includes lower-limb rehabilitation robotics data across task and acquisition variability, including fatigue, inter-day differences, and shift conditions [8]. When these degraded patterns enter the calibration set, the adaptation gradient follows the polluted calibration distribution. If every window is trusted equally, the prediction head can move toward a boundary that fits the corrupted session but performs worse on later clean data. Sensor degradation can therefore become calibration degradation and, in turn, harmful adaptation.

Simulated degradation and real non-ideal acquisition stress different aspects of the same adaptation problem. BLISS evaluates the early gait-phase prediction protocol, whereas K2MUSE provides mechanism-level evidence for the reliability-controlled adaptation rule under real non-ideal calibration.

2.4. Reliability-Controlled Adaptation

Several neighboring families of methods stabilize transfer in different ways. Meta-learning changes the source-training objective, prototypical learning changes the classifier geometry, domain adaptation aligns source and target representations, parameter-efficient tuning restricts the updated parameter set, and source-centered regularization penalizes departure from pretrained parameters [26,27,29,32,33]. Reliability-aware test-time methods can also select informative samples and preserve important source weights during an unlabeled test stream [34]. DTR-HA addresses a different setting: labeled new-user calibration is available, and its estimated quality controls both sample contribution and source-head anchoring after source training.

The closest controlled comparisons in this study therefore keep the source model and calibration information fixed. Weight-only HA changes the sample contribution, Fixed-anchor HA applies a static source penalty, and DTR-HA combines reliability weighting with reliability-scaled anchoring. Low-learning-rate and full-network controls test whether the result follows simply from smaller steps or a larger adaptation surface. The numerical claims in this paper are limited to these evaluated adaptation-stage comparisons; we do not claim the superiority of our method over meta-learning, prototypical-learning, or domain-adaptation methods.

Table 1 summarizes the methodological position of DTR-HA relative to the adjacent adaptation families discussed above.

Table 1. Methodological position of DTR-HA relative to adjacent adaptation families. The comparison is conceptual; no numerical superiority over the external method families is claimed. Abbreviations: DTR-HA, Dynamic Trust-Region Head Adaptation; MAML, model-agnostic meta-learning.

Method Family	Main Intervention	Calibration-Quality Role	Relation to DTR-HA
MAML/meta-learning [26]	Learns an initialization for rapid task adaptation	Not required by the original objective	Can provide the source initialization before DTR-HA
Prototypical learning [27]	Learns an episodic metric-space classifier	Not required by the original objective	Changes the classifier formulation; could be combined with reliability control
Feature-level domain adaptation [32]	Aligns or modifies source/target representations	Targets domain discrepancy rather than calibration trust	Can precede a reliability-controlled head update
Parameter-efficient tuning [29]	Restricts trainable parameters	No direct calibration-quality control	Small update surface
Test-time adaptation with sample selection [34]	Selects test samples and preserves source weights	Reliability acts on an unlabeled test stream	Adjacent sample-selection setting
Source-centered regularization [33]	Penalizes departure from pretrained parameters	Static reference penalty	Closest anchoring relation
DTR-HA	Weights calibration samples and scales source anchoring	Window- and set-level calibration reliability	Controls labeled new-user head adaptation

3. Materials and Methods

3.1. Study Overview

Figure 1 summarizes the framework. A source temporal convolutional network (TCN) is trained in each LOSO fold. For a held-out user, complete calibration bouts are used for personalization and later disjoint bouts are used for final testing. During personalization, the temporal encoder is frozen and only the prediction head is updated. DTR-HA controls this update using calibration reliability.

Full experimental-configuration details are provided in Supplementary Table S7.

3.2. Role-Separated Public-Dataset Reliability Design

The experiments use public datasets in complementary roles. BLISS, provided by the University of Bath Research Data Archive (Bath, UK; Version 3) (DOI: <https://doi.org/10.15125/BATH-01425>), is the target task dataset: each fold trains on 20 subjects, holds out one unseen user, uses early complete bouts for calibration, and reserves later disjoint bouts for final 0/100/200 ms gait-phase prediction. K2MUSE, provided through its official project page (<https://k2muse.github.io/>) and described in the corresponding *IJRR* article (DOI: <https://doi.org/10.1177/02783649261444046>), supplies real non-ideal calibration variability by mixing held-out-user Ideal cycles with Fatigue, ShiftAll, or broader non-Ideal cycles before testing on disjoint Ideal cycles. In the K2MUSE stress test, the downstream task classifier predicts terrain class (level ground, stair ascent, or stair descent) from cycle-level features, while the reliability scorer estimates Ideal versus non-Ideal calibration context. As a secondary target-pipeline check, we corrupt only held-out BLISS calibration bouts while keeping source training and final test bouts clean. This check links the K2MUSE non-ideal calibration evidence back to the BLISS gait-prediction setting without treating it as a separate general protocol. This role-separated design prioritizes reproducibility, leakage control, simulated and real calibration stress tests, negative controls for reliability, and adaptation-stage drift analysis.

Reliability-controlled new-user adaptation for wearable gait prediction

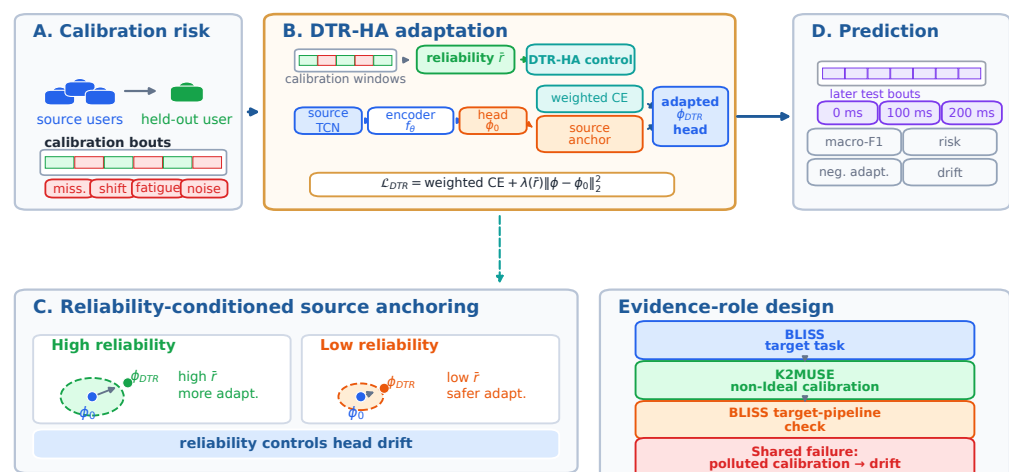


Figure 1. Main architecture of reliability-controlled new-user gait-prediction adaptation. Panel (A) shows short calibration bouts that may be clean or polluted. Panel (B) shows DTR-HA: reliability controls sample weighting and source-head anchoring while the encoder remains frozen. Panel (C) illustrates reliability-dependent source anchoring. Panel (D) summarizes disjoint 0/100/200 ms evaluation and outcome metrics. The evidence-role strip connects BLISS and K2MUSE evidence through calibration-induced head drift. Abbreviations: BLISS, Bilateral Lower-Limb Neuromechanical Signals dataset; CE, cross-entropy; DTR-HA, Dynamic Trust-Region Head Adaptation; TCN, temporal convolutional network.

3.3. Preprocessing

All preprocessing operations are fit only on the training subjects of the current LOSO fold. Non-finite values are converted to missing entries. Features that are entirely non-finite in the training fold are removed, and remaining missing values are imputed using training-fold medians. A signed logarithmic transform,

$$\tilde{x} = \text{sign}(x) \log(1 + |x|), \quad (1)$$

is applied before standardization. The training-fold imputer, feature mask, and scaler are then reused for held-out calibration and final test data. No held-out test statistics are used for preprocessing.

For BLISS, temporal windows contain five consecutive 50 ms frames. Horizon labels are generated by shifting the target within each complete bout; terminal samples without a future label are discarded. For K2MUSE, cycle-level feature rows are used for the real non-ideal calibration stress test.

3.4. Problem Formulation

Let $\mathbf{X}_t = [\mathbf{x}_{t-L+1}, \dots, \mathbf{x}_t] \in \mathbb{R}^{L \times C}$ be a window of wearable-device data ending at time t , where L is the sequence length and C is the feature dimension. For prediction horizon $\Delta_h \in \{0, 100, 200\}$ ms, the target is

$$y_t^{(h)} = y(t + \Delta_h). \quad (2)$$

In each LOSO fold, subject s is held out and the source model is trained on all other subjects. The held-out user's data are split by complete walking bouts into a calibration set $\mathcal{A}_s^{(k)}$ and a disjoint test set $\mathcal{T}_s$. No adjacent-window random split is used between calibration and final testing.

The calibration budgets for BLISS are $k \in \{0, 1, 3, 5, 10\}$ complete walking bouts. The $k = 0$ case is the source-only model. For $k > 0$, the source TCN encoder remains frozen and the prediction head is updated using only the selected calibration bouts. The final test set is unchanged across methods within each fold and horizon, enabling direct comparisons between the Source-only, Plain HA, Weight-only HA, Fixed-anchor HA, and DTR-HA methods. The budget is expressed in complete walking bouts rather than randomly sampled windows because this better approximates a short new-user setup session. The setting of 1–10 bouts represents a low-burden personalization regime: the user provides only a small number of complete walking trials before final evaluation on disjoint later bouts. No fixed duration is assumed because bout duration is dataset-dependent.

3.5. Source TCN and Head-Only Personalization

The source model is

$$\mathbf{p}_\theta(\mathbf{X}) = \text{softmax}(g_\phi(f_\theta(\mathbf{X}))), \quad (3)$$

where f_θ is a temporal convolutional encoder and g_ϕ is the prediction head. The BLISS TCN uses three 1-D convolutional blocks with kernel size 3 and dilations 1, 2, and 4, followed by a linear prediction head. For each LOSO fold, the source TCN is trained on the training subjects and then frozen during adaptation to the held-out user; the experiments do not reuse cached source checkpoints across folds.

The TCN is intentionally lightweight to keep the adaptation-stage comparison interpretable. A fixed, compact encoder reduces the chance that few-shot results are dominated by a large representation change and makes head-only and full-network controls easier to compare. In the main BLISS runs, the hidden dimension is 64, the dropout is 0.2, the

batch size is 512, and the source training window cap is 40,000 windows per fold/horizon. Adam-style optimization with weight decay is used. During personalization, only the linear head parameters are updated, keeping adaptation fast and auditable. This choice is compatible with wearable front ends, but it is made here for experimental control rather than as evidence of hardware deployment.

3.6. Plain, Weight-Only, and Fixed-Anchor Head Adaptation

Plain Head Adaptation updates the prediction head with all calibration samples equally using the cross-entropy (CE) loss:

$$\mathcal{L}_{\text{plain}}(\phi) = \frac{1}{N} \sum_{i=1}^N \text{CE}(y_i, g_{\phi}(f_{\theta}(\mathbf{X}_i))). \quad (4)$$

Weight-only Head Adaptation assigns reliability score $r_i \in [0, 1]$ to each calibration window and uses sample weight $w_i = s(r_i)$:

$$\mathcal{L}_{\text{weight}}(\phi) = \frac{1}{\sum_i w_i + \epsilon} \sum_i w_i \text{CE}(y_i, g_{\phi}(f_{\theta}(\mathbf{X}_i))). \quad (5)$$

Here and in Equation (7), ϵ is a small numerical constant set to 10^{-8} to avoid division by zero. It is used only as a numerical safeguard. The implementation applies a positive weight floor and normalizes nonzero weights over the calibration set before mini-batches are formed. Under this procedure, the denominator remained well away from zero in every nonempty adaptation batch used in the reported experiments. The zero-bout setting uses the Source-only model and does not evaluate the adaptation loss, while the positive floor prevents an all-zero weighting case. Thus, ϵ is not a tuned hyperparameter; changing it by several orders of magnitude around 10^{-8} would not materially affect the reported optimization. In the main experiments, $s(r_i) = \max(r_i, 10^{-6})^{\gamma}$ with $\gamma = 1$, followed by mean normalization of the nonzero weights over the calibration set before mini-batches are formed. Thus, lower-reliability calibration windows contribute less to the adaptation loss, while the normalization keeps the effective loss scale comparable. Fixed-anchor Head Adaptation adds a static source-head penalty:

$$\mathcal{L}_{\text{fixed}}(\phi) = \mathcal{L}_{\text{plain}}(\phi) + \lambda_0 \|\phi - \phi_0\|_2^2, \quad (6)$$

where ϕ_0 is the source prediction head. The notation $\|\cdot\|_2$ denotes the Euclidean norm over all prediction-head parameters, including the weight matrix and bias vector. The squared norm is used in the anchoring losses, whereas the unsquared norm is later reported as prediction-head drift.

These three baselines separate the components of the adaptation problem. Plain HA lets the calibration loss freely move the head. Weight-only HA changes which calibration samples influence the loss, but if the retained weighted evidence is still biased, the final head may drift too far. Fixed-anchor HA constrains drift, but the boundary is the same whether the calibration set is clean or polluted. The missing mechanism is therefore a reliability-dependent boundary on the final personalization step.

3.7. Dynamic Trust-Region Head Adaptation as Source Anchoring

DTR-HA combines reliability weighting with reliability-scaled source anchoring:

$$\mathcal{L}_{\text{DTR}}(\phi) = \left[\frac{1}{\sum_i w_i + \epsilon} \sum_{i=1}^N w_i \text{CE}(y_i, g_{\phi}(f_{\theta}(\mathbf{X}_i))) \right] + \lambda(\bar{r}) \|\phi - \phi_0\|_2^2. \quad (7)$$

The normalization term applies only to the reliability-weighted cross-entropy term; the source-head penalty is added as a separate anchoring term. Here, $\bar{r} = N^{-1} \sum_i r_i$ is the calibration-set reliability. The source-anchor strength is larger when calibration reliability is lower:

$$\lambda(\bar{r}) = \lambda_0[1 + \kappa(1 - \bar{r})]. \quad (8)$$

Thus, reliable calibration receives a weaker source penalty, whereas unreliable calibration receives a stronger penalty against movement from the source head. The temporal encoder remains fixed, and so this regularization acts directly on prediction-head personalization.

The main experiments use $\lambda_0 = 0.001$ and $\kappa = 1$. These values were retained from preliminary development and then fixed for the full-scale runs; they were not retuned by held-out subject or condition group, and no nested search was used to describe them as optimal. With $\kappa = 1$, the anchor coefficient varies smoothly from 0.001 at full reliability to 0.002 at zero reliability.

Here, “trust region” denotes a soft source-centered penalty on the prediction head rather than a hard constrained optimizer. The scalar κ controls how strongly the anchor tightens as reliability decreases. Within the reliability-weighted DTR-HA sensitivity family, $\kappa = 0$ gives a static-anchor reference; it is not identical to Fixed-anchor HA because the cross-entropy term remains reliability-weighted.

Figure 2 illustrates how calibration reliability changes the allowed displacement of the personalized prediction head from its source initialization.

Reliability-scaled source anchoring in prediction-head parameter space

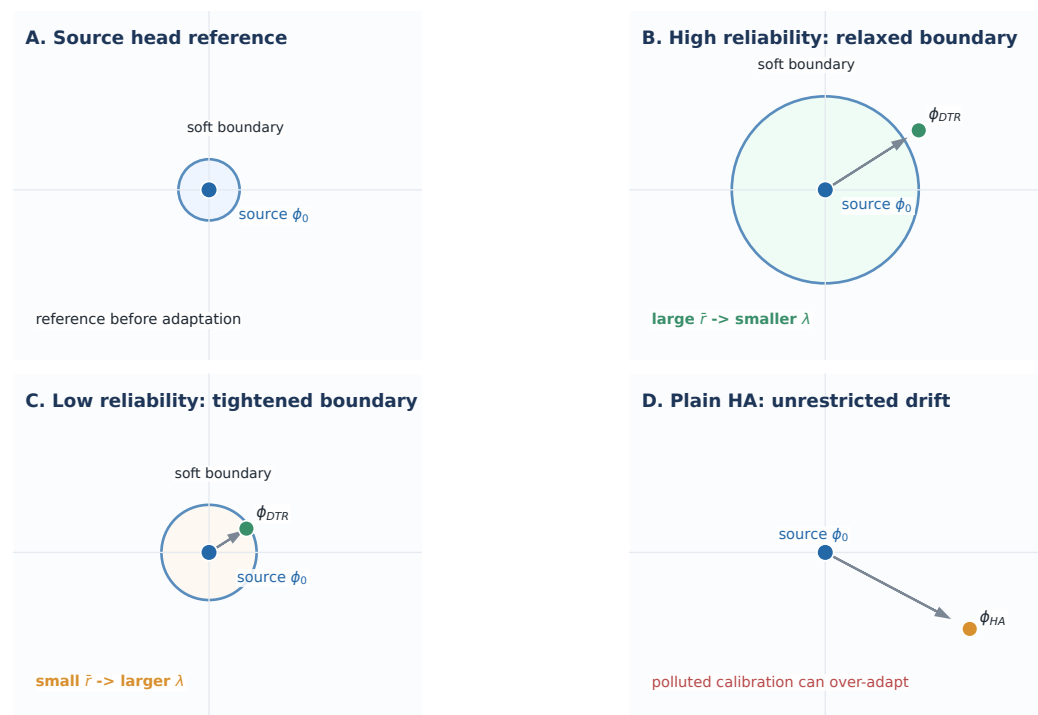


Figure 2. Reliability-scaled source anchoring in prediction-head parameter space. High calibration reliability relaxes the source-centered boundary and permits more personalization, whereas low reliability tightens the boundary and keeps the adapted head closer to ϕ_0 . Plain HA lacks this reliability-conditioned source boundary.

The practical random-forest (RF)-scored K2MUSE setting estimates reliability without held-out test labels. Oracle reliability is used only as an upper-bound diagnostic.

3.8. Reliability Scoring

Reliability is used here as an operational calibration-quality proxy rather than a direct physical measurement such as electrode impedance or a physiological EMG signal-quality index. The BLISS and K2MUSE scores are dataset-specific implementations of the same engineering variable: whether the current calibration evidence should be trusted strongly enough to drive new-user personalization.

In BLISS, reliability is a transparent sensing-availability and abnormality proxy based on finite-value ratio, missing-value rate, flat-channel indicators, and robust outlier indicators under simulated degradation. In K2MUSE, the held-out blind RF scorer is a context-based calibration-quality scorer: it distinguishes Ideal from non-Ideal acquisition-context distribution shifts using precomputed cycle-level features after excluding metadata identifiers. Training rows come only from non-held-out subjects, and the binary scorer label is Ideal versus non-Ideal condition. Held-out subject labels are not used to fit the scorer, choose DTR weights, or train the downstream terrain classifier. A fatigue-labeled cycle may still contain usable waveform segments, and so the RF score is interpreted as an acquisition-context proxy rather than an electrode-impedance measurement or physiological fatigue estimator. Optional selective prediction and risk-coverage curves are reported in Supplementary Figure S1.

3.9. Metrics

The primary predictive metric is macro-F1. For unreliable calibration, empirical risk, negative adaptation, subject-level wins, and prediction-head drift are used as stability-oriented engineering metrics. Risk is the empirical error rate, $1 - M^{-1} \sum_{j=1}^M \mathbf{1}[\hat{y}_j = y_j]$, computed on the held-out test windows or cycles. Because risk is computed from prediction errors, it is not necessarily equal to $1 - \text{macro-F1}$. Subject-level wins and losses are paired comparisons of macro-F1 against the stated baseline, usually Plain HA, for the same subject, condition group, and prediction horizon. Negative adaptation counts held-out subjects for which an adapted model has lower macro-F1 than the Source-only model on the same test set. We also report prediction-head drift,

$$d_{\text{head}} = \|\phi - \phi_0\|_2. \quad (9)$$

Selective-prediction coverage and accepted-risk definitions are provided in Supplementary Note S1 and illustrated in Supplementary Figure S1.

4. Results

4.1. Early Prediction and Few-Shot Adaptation Established on BLISS

Table 2 summarizes the 21-subject BLISS LOSO early prediction result. Source-only macro-F1 decreases from 0.5710 at the current phase to 0.5432 and 0.5096 at 100 and 200 ms horizons, confirming that future prediction is harder. This trend is expected because the model must infer the future phase from an observation window that may precede the next transition. The moderate 0 ms source-only macro-F1 should also be interpreted under the strict protocol used here: subjects are fully held out, calibration and test windows come from disjoint complete bouts, and no adjacent-window random split is used. Thus, the 0 ms setting is a strict unseen-user current-phase baseline rather than an easier within-subject random-window classification task. The result also confirms that the task is not a trivial current-state classification problem.

Few-shot head adaptation improves results for most held-out users. With 10 complete walking bouts, macro-F1 increases to 0.6116, 0.5788, and 0.5515 at 0, 100, and 200 ms, corresponding to gains of +0.0405, +0.0356, and +0.0420. Importantly, these gains are

obtained under complete-bout separation: calibration and final test windows do not come from the same walking bout.

Table 2. BLISS early gait-phase prediction and few-shot personalization. Values are mean macro-F1 $\pm$ standard deviation across 21 LOSO folds. Gains and improved-subject counts are relative to the source-only 0-bout model at the same horizon.

Horizon	Budget	Macro-F1	Gain vs. Source-Only	Improved Subjects
0 ms	0 bout	0.5710 $\pm$ 0.2420	–	–
0 ms	1 bout	0.5912 $\pm$ 0.2305	+0.0202	17/21
0 ms	3 bouts	0.5951 $\pm$ 0.2301	+0.0241	15/21
0 ms	5 bouts	0.6042 $\pm$ 0.2275	+0.0332	18/21
0 ms	10 bouts	0.6116 $\pm$ 0.2203	+0.0405	20/21
100 ms	0 bout	0.5432 $\pm$ 0.2279	–	–
100 ms	1 bout	0.5615 $\pm$ 0.2299	+0.0183	18/21
100 ms	3 bouts	0.5644 $\pm$ 0.2310	+0.0212	17/21
100 ms	5 bouts	0.5705 $\pm$ 0.2334	+0.0273	17/21
100 ms	10 bouts	0.5788 $\pm$ 0.2303	+0.0356	19/21
200 ms	0 bout	0.5096 $\pm$ 0.2175	–	–
200 ms	1 bout	0.5335 $\pm$ 0.2179	+0.0239	19/21
200 ms	3 bouts	0.5326 $\pm$ 0.2212	+0.0230	18/21
200 ms	5 bouts	0.5415 $\pm$ 0.2202	+0.0319	18/21
200 ms	10 bouts	0.5515 $\pm$ 0.2159	+0.0420	20/21

The standard deviations are large because held-out subjects differ substantially. This heterogeneity motivates personalization and reliability-aware adaptation. The few-shot gains appear across all three horizons and improve results for most subjects in the 10-bout setting, establishing BLISS as the target early-prediction protocol.

4.2. K2MUSE Provides the Real Non-Ideal Mechanism Evidence

The K2MUSE endpoint analysis focuses on the high-pollution setting where the robustness advantage is clearest. Under 75% bad calibration, DTR-HA reduced empirical risk by 20–23% relative to Plain HA, with macro-F1 gains of +0.018 to +0.022 and lower prediction-head drift. Full dose–response trends are presented in Supplementary Figure S6 and the corresponding source tables. Oracle reliability is retained only as an upper-bound diagnostic.

Within this role-separated design, K2MUSE serves as a real non-Ideal calibration stress test for the reliability-controlled adaptation mechanism rather than as a second gait-phase benchmark. The shared object of analysis is prediction-head movement induced by polluted calibration, and so K2MUSE macro-F1 values are interpreted only within this stress-test setting and are not compared directly with BLISS.

Table 3 reports the main 75% bad-calibration endpoint. Risk and stability provide the clearest robustness signal: empirical risk decreases by about 20.6%, 20.5%, and 23.0% in All non-Ideal, Fatigue, and ShiftAll, respectively; subject-level paired wins total 8/10, 10/10, and 7/8; and prediction-head drift is lower than that of Plain HA in all three groups. Macro-F1 also improves by +0.0204, +0.0184, and +0.0219 relative to Plain HA. DTR-HA has the largest relative risk reduction and drift reduction in all three non-Ideal calibration groups. All non-Ideal and Fatigue provide the strongest paired evidence (Wilcoxon $p = 0.0098$ and $p = 0.0020$). ShiftAll exhibited a directional improvement with 7/8 paired wins, although the paired test was not statistically significant under the current sample size ($p = 0.1484$). Subject-level endpoint rows are provided in Supplementary Table S2.

Table 3. K2MUSE main evidence at 75% bad calibration. Positive $\Delta F1$, risk reduction, and drift reduction are relative to Plain HA within the same condition group. DTR-HA combines reliability weighting with reliability-controlled source anchoring. Bold values identify the DTR-HA row in each condition group.

Condition Group	Method	Macro-F1	$\Delta F1$	Risk	Risk Red. (%)	Wins vs. Plain	Head Drift	Drift Red.
All non-ideal	Plain HA	0.8947	0.0000	0.1020	0.0	–	0.4509	0.0000
All non-ideal	Weight-only HA	0.9087	+0.0140	0.0876	14.1	10/10	0.4384	+0.0125
All non-ideal	Fixed-anchor HA	0.8949	+0.0002	0.1018	0.2	5/10	0.4507	+0.0002
All non-ideal	DTR-HA	0.9150	+0.0204	0.0810	20.6	8/10	0.4168	+0.0341
Fatigue	Plain HA	0.9034	0.0000	0.0929	0.0	–	0.4171	0.0000
Fatigue	Weight-only HA	0.9155	+0.0120	0.0806	13.2	10/10	0.4177	–0.0006
Fatigue	Fixed-anchor HA	0.9034	0.0000	0.0930	–0.1	5/10	0.4184	–0.0013
Fatigue	DTR-HA	0.9219	+0.0184	0.0739	20.5	10/10	0.4040	+0.0131
ShiftAll	Plain HA	0.8988	0.0000	0.0991	0.0	–	0.4300	0.0000
ShiftAll	Weight-only HA	0.9165	+0.0177	0.0812	18.1	8/8	0.4155	+0.0145
ShiftAll	Fixed-anchor HA	0.8983	–0.0005	0.0996	–0.5	3/8	0.4313	–0.0013
ShiftAll	DTR-HA	0.9207	+0.0219	0.0763	23.0	7/8	0.4003	+0.0297

The method comparison in Table 3 separates the components of the dynamic trust region. Weight-only HA helps under bad calibration by reducing the influence of unreliable samples. Fixed-anchor HA constrains drift with a static boundary. DTR-HA combines reliability-weighted evidence with a reliability-scaled source boundary, which explains the lower risk and drift in the strongest polluted-calibration settings. Weight-only HA produced many small subject-level improvements, whereas DTR-HA achieved lower empirical risk and smaller head drift. The dynamic source anchor therefore improves conservative risk and drift control instead of simply maximizing the count of small positive changes.

4.3. Head Drift and Source Anchoring

DTR-HA keeps the adapted prediction head closer to the source head under polluted calibration. At 75% bad calibration, the method's head-drift reduction is aligned with macro-F1 gain across K2MUSE condition groups (Table 3), consistent with the intended source-anchoring mechanism. Detailed drift plots are provided in the Supplementary Materials.

At the subject level, DTR-HA improves results for 8/10 subjects in All non-Ideal calibration, 10/10 in Fatigue, and 7/8 in ShiftAll. These wins are relevant because the target problem is reliable unseen-user personalization rather than only a pooled mean. Detailed subject-level gain distributions are provided in Supplementary Figure S5.

4.4. BLISS Target-Pipeline Calibration Stress Test

This additional BLISS target-pipeline stress test asks the same mechanism-level question inside the target early gait-phase pipeline: can polluted calibration induce harmful head drift, and can reliability-controlled anchoring constrain it? Source training subjects and held-out final test bouts remain clean; simulated corruption is applied only to held-out calibration bouts.

At 75% bad calibration, DTR-HA keeps macro-F1 nearly unchanged relative to Plain HA (–0.0004, –0.0006, and –0.0029 at 0, 100, and 200 ms) while reducing the head-drift norm by +0.0105, +0.0080, and +0.0068. In this clean-source/polluted-calibration/clean-test target-pipeline stress-test setting, the main effect is performance-preserving drift control inside the gait-phase pipeline. Paired tests show drift-control effects at 0 and 200 ms, with a directional trend at 100 ms (0 ms: $p = 0.0175$; 100 ms: $p = 0.0646$; 200 ms: $p = 0.0038$).

Table 4 adds the Source-only reference. Plain HA and DTR-HA remain above Source-only macro-F1 for all three horizons, indicating that limited personalization remains useful.

The design question is how much head movement should be allowed when calibration is polluted.

Table 4. BLISS target-pipeline 75% bad-calibration trade-off. The main effect is performance-preserving drift control: DTR-HA keeps macro-F1 within 0.003 of Plain HA while reducing head drift, with reduced negative adaptation at 200 ms.

Horizon (ms)	Source-Only Macro-F1	Plain HA Macro-F1	DTR-HA Macro-F1	$\Delta F1$ vs. Plain	Head-Drift Reduction	Negative Adaptation Plain→DTR
0	0.5655	0.6029	0.6025	−0.0004	+0.0105	1/21→1/21
100	0.5261	0.5658	0.5652	−0.0006	+0.0080	2/21→3/21
200	0.5146	0.5486	0.5458	−0.0029	+0.0068	6/21→4/21

The compact trade-off table emphasizes that DTR-HA primarily controls drift while preserving macro-F1 within a small tolerance.

Supplementary subject-level analyses show horizon dependence. At 200 ms, DTR-HA reduced negative-adaptation cases from 6/21 to 4/21 and rescued two held-out subjects that Plain HA moved below the source-only model. The rescue effect was horizon-dependent: DTR-HA reduced negative adaptation at 200 ms, showed no change at 0 ms, and did not reduce the negative-adaptation count at 100 ms. This suggests that conservative anchoring may limit adaptation for a small number of heterogeneous users at intermediate horizons.

Additional BLISS diagnostics keep the analysis tied to gait prediction: near-transition windows are harder than steady-phase windows at future horizons, and head-only adaptation remains competitive under mixed-corruption calibration.

Detailed BLISS target-pipeline stress-test drift plots are provided in the Supplementary Materials.

4.5. Reliability Negative Controls and Adaptation-Stage Robustness

The reliability score is a lightweight sensing-availability and calibration-context proxy. Its role is to identify gross reliability changes relevant to adaptation, including missing channels, abnormal finite ratios, outliers, and held-out blind non-Ideal acquisition probabilities.

Reliability-ranked BLISS calibration, K2MUSE dose–response behavior, drift reduction, and supplementary correlation analyses all point to task-relevant calibration-quality information in the reliability score.

K2MUSE reliability negative controls compare normal RF-scored reliability with shuffled, random, inverse, and constant assignments while recomputing the adapted heads. These controls separate generic conservative protection from reliability-ordered adaptation. Table 5 reports normal reliability performance and the macro-F1 drops caused by shuffled, constant, inverse, or random reliability assignments, supporting the importance of meaningful reliability ordering. Thus, the strongest adaptation benefit depends on meaningful calibration-quality ordering, not only on adding a source penalty. These reruns test reliability-ordering effects and are not intended as a row-by-row reproduction of the Table 3 endpoint.

Rescue-harm analysis gives the same risk-oriented picture. Under 75% bad calibration, DTR-HA reduced negative adaptation from 4/10 to 0/10 in All non-Ideal calibration, from 1/10 to 0/10 in Fatigue, and from 3/8 to 1/8 in ShiftAll. In the BLISS target-pipeline test at 200 ms, negative adaptation decreased from 6/21 to 4/21, with two rescued held-out subjects.

Compact rescue-harm and evidence-closure summaries are provided in Supplementary Tables S3 and S6.

Table 5. K2MUSE reliability negative controls under 75% bad calibration. Normal RF macro-F1 is reported within the negative-control rerun protocol; the other columns show Δ macro-F1 relative to the normal reliability assignment. The table tests reliability-ordering effects rather than reproducing Table 3 endpoint.

Condition Group	Normal RF Macro-F1	Shuffled Δ F1	Constant Δ F1	Inverse Δ F1	Random Δ F1
All non-Ideal	0.9205	−0.0252	−0.0156	−0.0239	−0.0164
Fatigue	0.9220	−0.0200	−0.0125	−0.0231	−0.0141
ShiftAll	0.9351	−0.0251	−0.0198	−0.0255	−0.0234

In a dedicated 75% bad-calibration sensitivity rerun, the mean gain over Plain HA remained positive for every tested $\kappa \in \{0, 0.5, 1, 2, 4, 8\}$ in all three K2MUSE groups (Supplementary Figure S2). Lowering the Plain HA learning rate also failed to reproduce the combined macro-F1 and drift pattern: its best values were 0.8930, 0.8965, and 0.9036, compared with 0.9150, 0.9219, and 0.9207 for DTR-HA (Supplementary Table S4). The BLISS full-network sanity check likewise showed no systematic advantage over head-only adaptation across the three horizons (Supplementary Table S5).

5. Discussion

Across BLISS, K2MUSE, and the additional BLISS calibration-pollution check, the failure mode appears at the adaptation stage. A source model may be useful for gait prediction, yet a short calibration session can still move the new-user head in the wrong direction. K2MUSE is used here as a real non-ideal calibration stress test, not as a second gait-phase benchmark; the BLISS check asks whether the same drift behavior is visible inside the target pipeline.

The key methodological distinction is where reliability enters the update. Weight-only HA changes the contribution of individual windows, whereas Fixed-anchor HA applies the same source penalty to every calibration set. DTR-HA combines the two levels: window scores weight the calibration loss, and set-level reliability scales the penalty on whole-head displacement. The negative controls sharpen this distinction. Constant reliability provides some generic protection, but shuffled, random, and inverse assignments weaken the result, showing that the ordering of calibration quality matters.

This is adaptation-stage robustness rather than robustness to noisy test inputs or closed-loop robust control. The corrupted object is the model update itself. Negative adaptation is therefore a useful failure measure: it records cases in which personalization performs worse than retaining the Source-only head. Lower empirical error, fewer negative-adaptation cases, and smaller head drift reflect an update that is less harmful under polluted calibration; macro-F1 remains the primary predictive metric.

The practical significance is clearest in the high-pollution setting. K2MUSE shows a 20–23% relative reduction in empirical test error and wins in 25 of 28 paired subject-condition comparisons. The BLISS result is more mixed: macro-F1 remains within 0.003 of Plain HA, head drift decreases at every horizon, and the negative-adaptation count improves only at 200 ms. These results favor conservative personalization when calibration is doubtful, while leaving room for stronger user-specific adaptation when the evidence is reliable.

Head-only adaptation is deliberate in this setting. Freezing the temporal encoder preserves the gait representation learned from source users, reduces the number of parameters exposed to few-shot overfitting, and keeps the update easier to inspect. In the BLISS pipeline, only the 325-parameter head is updated. The same reliability-control idea may be useful for inertial-measurement-unit (IMU)-, electroencephalography (EEG)-, or electrocardiography (ECG)-based personalization, but each modality would require its

own failure model, reliability estimator, and validation; the present experiments do not establish those extensions.

Practical Deployment Considerations

DTR-HA requires a reliability estimate suited to the device and protocol, not a universal scorer. Historical labeled non-ideal data can support a held-out blind scorer such as the K2MUSE RF model; otherwise, missingness, finite-value ratio, flat-channel indicators, and robust outlier statistics provide transparent first-line proxies, as in the BLISS check. Very low-trust calibration may be handled by retaining the source head or requesting recalibration. In the logged BLISS runs, DTR-HA added about 2.6% adaptation-stage overhead and no extra prediction-head parameters after adaptation.

6. Limitations and Future Work

This is an offline public-dataset evaluation of reliability-controlled new-user adaptation. It does not include hardware-in-the-loop testing, closed-loop assistive-device control, assistive-force generation, or long-term online adaptation. The reliability estimators are intentionally lightweight and dataset-specific. BLISS uses a transparent sensing proxy, whereas the held-out blind K2MUSE RF scorer relies on historical Ideal versus non-Ideal acquisition-context labels. A new device without such labeled history would require transparent signal-quality proxies or unsupervised anomaly detection. The BLISS target-pipeline stress test is an auxiliary within-dataset check rather than evidence of cross-target generality. K2MUSE's subject-level sample size is limited, especially for ShiftAll, where only eight held-out subjects were available; ShiftAll showed a directional improvement with 7/8 paired wins, but the paired test was not statistically significant at the current sample size ($p = 0.1484$). The study evaluates adaptation behavior offline and does not test downstream control decisions.

Future work can examine reliability estimators that incorporate uncertainty, signal-quality measures, or conformal scores, together with robust adaptation losses such as generalized cross entropy or worst-group objectives. DTR-HA could also be applied after a meta-learned initialization or a domain-adapted representation instead of treating these approaches as substitutes [26,27,32]. Hardware-in-the-loop and closed-loop studies are still needed to determine whether the offline reduction in harmful personalization improves assistive-device timing.

7. Conclusions

This paper studies reliability-controlled new-user adaptation for lower-limb gait prediction with wearable devices under unreliable calibration. BLISS establishes the unseen-user early prediction protocol, K2MUSE provides real non-ideal calibration evidence, and an additional BLISS calibration-pollution test checks the same drift behavior inside the target pipeline. DTR-HA combines reliability-weighted calibration loss with a reliability-scaled soft penalty around the source head. Under polluted calibration, it reduces empirical risk, harmful prediction-head drift, and negative adaptation in the K2MUSE stress tests while keeping the BLISS target-pipeline macro-F1 within a small tolerance and reducing head drift. The resulting design rule is direct: calibration quality should determine how strongly a new-user update is anchored to the source head.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/app16157661/s1>. The supplementary material is supplied as one PDF and contains: Table S1, BLISS early-prediction and few-shot adaptation; Table S2, K2MUSE subject-level DTR-HA gains; Table S3, BLISS target-pipeline stress-test trade-off summary; Table S4, low-learning-rate Plain HA controls; Table S5, BLISS transition-window, per-phase, and full-network

fine-tuning analyses; Table S6, reliability and robustness controls; Table S7, experimental configuration checklist; Figure S1, selective-prediction risk–coverage curve; Figure S2, source-anchor sensitivity; Figure S3, reliability-score audit; Figure S4, BLISS transition-window diagnostics; Figure S5, K2MUSE subject-level DTR-HA gains; Figure S6, K2MUSE bad-calibration dose–response; and Supplementary Note S1, selective-prediction risk–coverage definitions.

Author Contributions: Conceptualization, Z.Z. and K.L.; methodology, Z.Z.; software, Z.Z.; validation, Z.Z.; formal analysis, Z.Z.; investigation, Z.Z.; data curation, Z.Z.; writing—original draft preparation, Z.Z.; writing—review and editing, Z.Z. and K.L.; visualization, Z.Z.; supervision, K.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The experiments use publicly available datasets. BLISS is available through the University of Bath Research Data Archive at the following <https://doi.org/10.15125/BATH-01425>. K2MUSE is described in the *IJRR* article at the following <https://doi.org/10.1177/02783649261444046>. K2MUSE is distributed through its official project and Kaggle dataset pages. Data use is subject to the licenses and terms specified by the corresponding dataset providers.

Acknowledgments: The authors acknowledge the public release of the BLISS and K2MUSE datasets used in this study. The OpenAI Codex desktop application (GPT-5.3-Codex-Spark and GPT-5.5; accessed June–July 2026) was used only for language polishing, formatting assistance, and manuscript organization support. All scientific content, analyses, experimental results, interpretations, and final decisions were reviewed and approved by the authors, who take full responsibility for the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Huang, H.; Zhang, F.; Hargrove, L.J.; Dou, Z.; Rogers, D.R.; Englehart, K.B. Continuous Locomotion-Mode Identification for Prosthetic Legs Based on Neuromuscular-Mechanical Fusion. *IEEE Trans. Biomed. Eng.* **2011**, *58*, 2867–2875. [\[CrossRef\]](#) [\[PubMed\]](#)
- Young, A.J.; Simon, A.M.; Hargrove, L.J. Intent Recognition in a Powered Lower Limb Prosthesis Using Time History Information. *Ann. Biomed. Eng.* **2014**, *42*, 631–641. [\[PubMed\]](#)
- Young, A.J.; Kuiken, T.A.; Hargrove, L.J. Analysis of using EMG and mechanical sensors to enhance intent recognition in powered lower limb prostheses. *J. Neural Eng.* **2014**, *11*, 056021. [\[CrossRef\]](#) [\[PubMed\]](#)
- Tucker, M.R.; Olivier, J.; Pagel, A.; Bleuler, H.; Bouri, M.; Lambercy, O.; del R. Millan, J.; Riener, R.; Vallery, H.; Gassert, R. Control Strategies for Active Lower Extremity Prosthetics and Orthotics: A Review. *J. Neuroeng. Rehabil.* **2015**, *12*, 1. [\[CrossRef\]](#) [\[PubMed\]](#)
- Prasanth, H.; Caban, M.; Keller, U.; Courtine, G.; Ijspeert, A.; Vallery, H.; von Zitzewitz, J. Wearable Sensor-Based Real-Time Gait Detection: A Systematic Review. *Sensors* **2021**, *21*, 2727. [\[CrossRef\]](#) [\[PubMed\]](#)
- Geng, Y.; Zhou, P.; Li, G. Toward attenuating the impact of arm positions on electromyography pattern-recognition based motion classification in transradial amputees. *J. Neuroeng. Rehabil.* **2012**, *9*, 74. [\[CrossRef\]](#) [\[PubMed\]](#)
- Phinyomark, A.; Scheme, E. EMG pattern recognition in the era of big data and deep learning. *Big Data Cogn. Comput.* **2018**, *2*, 21. [\[CrossRef\]](#)
- Li, J.; Zhang, B.; Tan, X.; Chen, W.; Liu, Z.; Zhang, J.; Huo, W.; Huang, J.; Liu, L.; Zhao, X. K2MUSE: A human lower-limb multimodal walking dataset spanning task and acquisition variability for rehabilitation robotics. *Int. J. Robot. Res.* **2026**, OnlineFirst. [\[CrossRef\]](#)
- Ahmed, S.; Mohanna, M.; Martinez Hernandez, U.; Awad, M.; Mansour, A. *Datasets for Bilateral Lower-Limb Neuromechanical Signals in Able-Bodied and Impaired Individuals with Wearable and Ambient Sensors (BLISS)*; Version 3; University of Bath Research Data Archive: Bath, UK, 2025. [\[CrossRef\]](#)
- Camargo, J.; Ramanathan, A.; Flanagan, W.; Young, A. A Comprehensive, Open-Source Dataset of Lower Limb Biomechanics in Multiple Conditions of Stairs, Ramps, and Level-Ground Ambulation and Transitions. *J. Biomech.* **2021**, *119*, 110320. [\[CrossRef\]](#)
- Lencioni, T.; Carpinella, I.; Rabuffetti, M.; Marzegan, A.; Ferrarin, M. Human kinematic, kinetic and EMG data during different walking and stair ascending and descending tasks. *Sci. Data* **2019**, *6*, 309. [\[CrossRef\]](#) [\[PubMed\]](#)
- Moreira, L.; Figueiredo, J.; Santos, C.P.; Moreno, J.C. Human lower-limb kinematics, kinetics, and EMG data during locomotion on slopes and stairs. *Sci. Data* **2021**, *8*, 233.

13. Saboor, A.; Kask, T.; Kuusik, A.; Alam, M.M.; Le Moullec, Y.; Niazi, I.H.; Zoha, A.; Ahmad, R. Latest Research Trends in Gait Analysis Using Wearable Sensors and Machine Learning: A Systematic Review. *IEEE Access* **2020**, *8*, 167830–167864. [\[CrossRef\]](#)
14. Torricelli, D.; Rodriguez-Guerrero, C.; Veneman, J.F.; Crea, S.; Briem, K.; Lenggenhager, B.; Beckerle, P. Benchmarking Wearable Robots: Challenges and Recommendations From Functional, User Experience, and Methodological Perspectives. *Front. Robot. AI* **2020**, *7*, 561774. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Wang, T.; Li, Y.; Liu, J.; Peng, Q.; Su, H. Human Gait Analysis Using Wearable Sensors with Kinematic and Kinetic Algorithms. *Sensors* **2023**, *23*, 614. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Kang, I.; Molinaro, D.D.; Duggal, S.; Chen, Y.; Kunapuli, P.; Young, A.J. Real-Time Gait Phase Estimation for Robotic Hip Exoskeleton Control During Multimodal Locomotion. *IEEE Robot. Autom. Lett.* **2021**, *6*, 3491–3497. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Liu, Y.; An, H.; Ma, H.; Wei, Q. Online Walking Speed Estimation Based on Gait Phase and Kinematic Model for Intelligent Lower-Limb Prosthesis. *Appl. Sci.* **2023**, *13*, 1893. [\[CrossRef\]](#)
18. Xiong, B.; Lou, J.; Ni, W.; Su, Z.; Huang, S. A Multi-Branch Adaptive Model with Hybrid Time–Frequency Loss to the Enhanced Joint Moment Prediction of Prosthetic Control and Human Motion Applications. *Appl. Sci.* **2025**, *15*, 1678. [\[CrossRef\]](#)
19. Choi, W.; Yang, W.; Na, J.; Lee, G.; Nam, W. Feature Optimization for Gait Phase Estimation with a Genetic Algorithm and Bayesian Optimization. *Appl. Sci.* **2021**, *11*, 8940. [\[CrossRef\]](#)
20. Bai, S.; Kolter, J.Z.; Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv* **2018**, arXiv:1803.01271.
21. Lim, B.; Zohren, S. Time-Series Forecasting with Deep Learning: A Survey. *Philos. Trans. R. Soc. A* **2021**, *379*, 20200209. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Wang, J.; Chen, Y.; Hao, S.; Peng, X.; Hu, L. Deep Learning for Sensor-based Activity Recognition: A Survey. *Pattern Recognit. Lett.* **2019**, *119*, 3–11. [\[CrossRef\]](#)
23. Zhang, J.; Fiers, P.; Witte, K.A.; Jackson, R.W.; Poggensee, K.L.; Atkeson, C.G.; Collins, S.H. Human-in-the-loop optimization of exoskeleton assistance during walking. *Science* **2017**, *356*, 1280–1284. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Slade, P.; Kochenderfer, M.J.; Delp, S.L.; Collins, S.H. Personalizing Exoskeleton Assistance while Walking in the Real World. *Nature* **2022**, *610*, 277–282. [\[CrossRef\]](#) [\[PubMed\]](#)
25. McKean, M.; Wang, X.; Xiao, Y.; Wu, B.; Wang, Z.; Liu, Y.; Wang, P.; Chen, C.; Wu, X. Instantly Personalized Predictive Locomotion Mode Recognition Using Neural Mechanical Encoding. *Front. Neurosci.* **2024**, *18*, 1457623. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Finn, C.; Abbeel, P.; Levine, S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2017; Volume 70, pp. 1126–1135.
27. Snell, J.; Swersky, K.; Zemel, R. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4080–4090.
28. Perez, E.; Strub, F.; de Vries, H.; Dumoulin, V.; Courville, A. FiLM: Visual Reasoning with a General Conditioning Layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA, 2018; Volume 32. [\[CrossRef\]](#)
29. Hounsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; de Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2019; Volume 97, pp. 2790–2799.
30. Farago, P.; Groza, V.; Ivanciu, L.N.; Hintea, S.; Popa, V. Surface Electromyography for Muscle Activity and Motion Detection: A Review on Signal Quality and Processing. *Sensors* **2023**, *23*, 8860. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Sul, H.; Kim, H.S.; Kim, Y.; Hur, P.; Kim, H.M. Analysis of Signal Processing Methods for EMG-Based Recognition Systems in Lower Limb Exoskeletons. *Sensors* **2025**, *25*, 3459. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-Adversarial Training of Neural Networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.
33. Li, X.; Grandvalet, Y.; Davoine, F. Explicit Inductive Bias for Transfer Learning with Convolutional Networks. In *Proceedings of the 35th International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2018; Volume 80, pp. 2825–2834.
34. Niu, S.; Wu, J.; Zhang, Y.; Chen, Y.; Zheng, S.; Zhao, P.; Tan, M. Efficient Test-Time Model Adaptation without Forgetting. In *Proceedings of the 39th International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2022; Volume 162, pp. 16888–16905.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.